Mitigating Risk of Data Loss in Preservation Environments

Reagan W. Moore
San Diego Supercomputer Center
moore@sdsc.edu

Joseph F. JaJa
University of Maryland
joseph@umiacs.umd.edu

Robert Chadduck
National Archives and Records Administration
robert.chadduck@nara.gov

Abstract[1]

Preservation environments manage digital records for time periods that are much longer than that of a single vendor product. A primary requirement is the preservation of the authenticity and integrity of the digital records while simultaneously minimizing the cost of long-term storage, as the data is migrated onto successive generations of technology. The emergence of low-cost storage hardware has made it possible to implement innovative software systems that minimize risk of data loss and preserve authenticity and integrity. This paper describes software mechanisms in use in current persistent archives and presents an example based upon the NARA research prototype persistent archive.

1. Introduction.

Preservation environments, called persistent archives, support the long-term storage of digital records [1]. A persistent archive manages retention of both the digital record content and a preservation context that describes the origin, relevance, and preservation properties associated with each digital record. Persistent archives assert that the digital record remains authentic, that the digital record is a true copy of the original digital record deposited into the archive, and that the person and judicial processes that created the digital record remain correctly identified. Authenticity requires multiple preservation properties: that the digital record remains unchanged, that the preservation context correctly tracks information about preservation processes performed upon the digital record, and that the chain of custody of the digital record remains unbroken. A persistent archive provides the mechanisms to validate assertions of authenticity, even while the technology used to implement the persistent archive evolves over time [2]. For the National Archives and Records Administration (NARA), a reasonable goal for the preservation period is the lifetime of the republic, nominally 400 years. During this time period, the persistent archive will need to migrate the preserved content across one hundred generations of storage systems. In effect, the preservation environment needs to manage the preserved material independently of the choice of current storage technology.

The traditional approach to data management is to assume that the selected storage repository is responsible for maintaining authenticity and integrity of the deposited records. Since current file systems do not manage the metadata attributes required for asserting authenticity, the preservation metadata can be packaged with each digital entity in an Archival Information

[1] The Storage Resource Broker was developed under the technical lead of Michael Wan and Arcot Rajasekar at the San Diego Supercomputer Center. Applications of the SRB technology were done by Wayne Schroeder, George Kremenek, Sheau-Yen Chen, Charles Cowart, Lucas Gilbert, Bing Zhu, and Marcio Faerman. This work was supported in part by the NSF NPACI ACI-9619020 (NARA supplement), the NSF Digital Library Initiative Phase II Interlib project, the NSF NSDL/UCAR Subaward S02-36645, the DOE SciDAC/SDM DE-FC02-01ER25486 and DOE Particle Physics Data Grid, the NSF National Virtual Observatory, the NSF Grid Physics Network, the NSF Southern California Earthquake Center, and the NASA Information Power Grid. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archives and Records Administration, or the U.S. government.

0-7695-2318-8/05 $20.00 © 2005 IEEE

Package (AIP) [3]. The AIP is written to a storage repository, and a separate database is used to track the location of each electronic record. The requirements for preservation of authenticity can then be imposed as hardware capabilities that ensure against data corruption or data loss. The commercial storage system is tasked with replicating the data, mirroring the state information about the storage location of each AIP, maintaining a file name space for the AIP, maintaining information about archivists who are allowed to execute preservation processes, managing access controls, and managing audit trails to track what has happened. Such systems can implement the preservation requirements in hardware and incorporate software systems to manage the preservation context (typically a database). The archivist picks the solution that minimizes risk of loss of authenticity, while simultaneously minimizing total cost of preservation.

With the advent of low-cost commodity storage systems [4], it becomes possible to lower the cost of long-term preservation by moving risk mitigation technology into the preservation environment software. Instead of requiring that the storage system internally protect against data loss through hardware redundancy, the electronic records can be replicated onto multiple lower-cost storage systems. This paper looks at the types of software mechanisms that help mitigate the risk of data loss, examines how the software mechanisms are implemented in data grids, and provides an example of a preservation system based upon the NARA research prototype persistent archive. Special attention is paid to the scalability of the preservation software mechanisms, to ensure that preservation environments can sustain the petabytes of data and millions of records that are now being created.

2. Types of risk for loss of authenticity

One way to proceed with a risk analysis is to quantify the types of data and information that must be preserved. We characterize the bits that comprise the digital records as the *content*, and the preservation metadata for asserting authenticity as the *context*. We can then consider systems that optimize the preservation of content (such as file systems and archival storage systems), systems that optimize the preservation of context (such as databases), and the systems that are needed to associate context with content (such as data grids). This approach makes it possible to implement management mechanisms that are scalable. Content is streamed through preservation procedures in bulk operations and stored on storage repositories after aggregation in containers. Context is managed in databases using bulk metadata manipulation procedures on metadata that has been aggregated into XML files.

The assertion of authenticity requires that all operations performed upon a digital entity can be tracked, and that the resulting state information can be maintained for examination at any future time. Data grids provide this capability. The preservation metadata is maintained in a database along with a handle that points to the storage location of the electronic record. The preservation context is archived as files through dumps of the database metadata. Thus both the preservation metadata and the electronic records can be encapsulated in files. If the integrity of the electronic records and the integrity of the database can be maintained, then the authenticity can also be maintained if the preservation context is updated after each operation on a preserved electronic record.

The mechanisms that ensure scalability must interoperate with the mechanisms that protect against loss of authenticity. We consider software systems that support bulk operations on replicated content and federated context. Replication of content corresponds to the creation of multiple copies and the tracking of each copy. Federation of context corresponds to synchronization of two independent databases (preferably from different vendors) that each hold the preservation metadata under constraints imposed for both access controls and update consistency [5].

Typical risks and the associated software risk mitigation mechanisms are:

- Media failure – handled by replication on multiple media. Examples of risks include disk crashes and tape corruption. At the San Diego Supercomputer Center, current commodity disks have a mean lifetime of 6 years. A 15-Terabyte disk farm made of 200-GB disks has a disk failure on average once a month. Tape lifetimes are similar (on the order of five to ten years). Both disks and tapes are replaced predominantly to recover floor space by using higher capacity media. Disk and tape errors typically compromise files in a storage volume.
- Vendor-hardware/software systemic error – handled by replication onto another vendor product. Examples of risks include corruption caused by RAID controllers when writing data and loss of location information through database corruption. Such problems can compromise all electronic records on the corrupted system.
- Operational error – handled by federation with an independent preservation environment. Examples of risks include procedures that are compromised during upgrades to new storage systems. The assumptions under which a procedure is executed may no longer be valid for the new storage environment. Such problems can compromise all electronic records written after the system upgrade.
- Natural disaster –handled by federation with a preservation environment at a geographically remote site. Examples of risks include fire, flood, and earthquake. While quite infrequent, the entire archive may be compromised.
- Malicious user – handled by federation with a deep archive. Examples of risks include compromise of the Unix operating system environment through security holes that are not related to the storage system. The authenticity of electronic records cannot be assured when either the preservation context or the content may be changed surreptitiously. A deep archive is a preservation environment that forces all accesses to be local, that handles all updates through creation of versions, and that prohibits external user access. Content and context are sent to a staging area for ingestion into the deep archive. Data and metadata are moved from the staging area under the control of archivist procedures.

Protecting against all types of risk requires the distribution of data across multiple sites using multiple vendor products. A minimum of two sites is required with the second site acting as a slave to the first site. Data and metadata that are registered into the first site are asynchronously sent to the second site. The second site can be implemented at a geographically remote location using storage systems and databases provided by different commercial vendors than used for the first site. The second site storage systems can be operated independently of the first site, with upgrades to new equipment and procedures done at different times. Each site supports a separate preservation metadata database.

The deep archive can be implemented as a separate data management system at one of the sites using an independent database and storage repository. Since both the preservation metadata and electronic records must be staged into the deep archive, the deep archive can also be implemented at a third geographic location. The goal is to minimize the opportunity for a single person to compromise all copies of a record and the associated preservation metadata.

The distribution of replicas and metadata across multiple sites using heterogeneous storage systems managed under separate administrative domains requires the use of data grid technology. Data grids provide the interoperability mechanisms needed to manage data distributed across multiple types of storage systems. Federations of data grids provide the authenticity coordination needed to replicate preservation metadata. By replicating both preservation context and the preservation content onto multiple systems, it is possible to decrease the reliability requirements for each individual system, while providing an opportunity to address risks that are inherent in relying upon technology from a single vendor.

3. Minimizing cost of preservation

Minimization of preservation cost is achieved by relying upon commodity storage platforms. The San Diego Supercomputer Center uses Grid Bricks [4] to provide low-cost commodity-based disk storage for multi-terabyte collections. The Grid Bricks are modular storage systems that integrate a 1.7 Ghz CPU, a gigabyte of memory, a Gigabit Ethernet network connection, and 5 Terabytes of disk. At the end of calendar year 2004, such systems could be implemented at a cost of $2000 per terabyte.

Each Grid Brick provides a minimal set of capabilities:

- naming convention for files (physical file name)
- naming convention for users (file owner)
- storage location (network address)
- association of a context with each digital entity (typically creation time, file size, ownership, update time)
- consistency controls for the update of the context and access controls (implemented through a Linux file system)

A preservation environment must extend these capabilities across multiple Grid Bricks, across distributed storage sites and between federated data grids. This is equivalent to implementing a name space for each capability that is controlled and managed by the preservation environment, independently of the commodity storage system. The preservation environment needs to own and manage five name spaces [6]:

- naming convention for the digital entities (logical file name)
- naming convention for the users (distinguished user names)
- naming convention for the storage resources (logical resource name)
- naming convention for the context attributes (metadata name space)
- naming convention for the consistency constraints (knowledge concept space)

Each naming convention implements a logical name space for the associated identifiers. The logical name space for digital entities is taken as the principal name space onto which the preservation context, access control and consistency constraints are mapped [7]. The persistent archive maps from the logical file name space to the physical file name space as provided by a particular vendor product (storage repository file name, database binary large object, object in a ring buffer). The distinguished name space for users provides a single sign-on authentication environment, using systems such as Grid Security Infrastructure certificates or a challenge response authentication mechanism. The user names are managed independently of the storage systems by having each electronic record written under a Unix ID associated with the data grid. Access controls are maintained by the data grid on the logical file name space. A user authenticates himself/herself to the data grid, the data grid checks the access controls for permission to manipulate the electronic record, the data grid then authenticates itself to the storage repository, and then the file is retrieved and transmitted to the user.

The logical resource name space makes it possible to associate operations with sets of physical resources. Replication can be implemented as a write operation on a logical resource name with the constraint that the write completes when copies exist on each of the storage systems in the list. Load leveling can be implemented as a write operation on a logical resource name with the constraint that the write completes when a copy exists on the next storage system in the list. Similarly, fault tolerance can be implemented as a constraint that a write completes when copies have been made on “k” of the “n” storage resources in the list.

The consistency constraints define relationships that are imposed on interactions between the name spaces. Examples include:

- Access control, which is a constraint imposed between the user name space, the logical file name space, and allowed operations on the digital entities,
- Authenticity, which is a constraint on the update of state information associated with each electronic record (location, audit trails, links to replicas) based on the role of the person issuing an operation request,
- Integrity, which is a constraint imposed on the frequency with which checksums are evaluated for each electronic record and compared to the preservation metadata value.

Data grids implement the above logical name spaces and consistency constraints [8]. By integrating multiple commodity-based storage systems into a data grid, the creation and management of replicas of electronic records can be automated. Data grids provide the control mechanisms needed to maintain consistency between the multiple copies. By federating multiple data grids, the preservation context can also be replicated, ensuring the preservation of authenticity metadata against the risk mechanisms listed above.

4. Scalability

Preservation environments now support tens of millions of files and petabytes of data. The ability to manage such massive collections requires automation of the management of the associated preservation context, the automation of the management of the preservation content, and the automation of data movement and metadata registration. Preservation environments are now feasible because commercial database systems are capable of managing billions of records. Preservation metadata for each electronic record can be effectively managed. Commercial file systems, however, do not scale as well. Most file systems are capable of managing ten to twenty million files before the file system i-node structure performance degrades. The use of commodity-disk Linux file systems appears to be a major bottleneck for scaling to a hundred million files. Data grids that manage data distributed across multiple commercial storage systems have to provide scalability mechanisms to overcome file system name space limitations.

Standard approaches to overcoming file system limitations are file distribution across multiple file systems (load leveling across Grid Bricks) and aggregation of files into containers before storage. Load leveling is managed by data grid collective operations on logical resource. Data grids also support file aggregation into containers. After aggregation, the file system sees only the containers, while the preservation database records the offset location of each electronic record within a container. Data grids support use of containers across all types of storage repositories, whether Unix file systems, Windows file systems, or archival storage systems. This makes it possible to implement scalable data management systems. However, the registration and loading of electronic records into the preservation environment can still present a major bottleneck.

Three different types of registration and loading need to be considered:

- From "local file system name space" to the "data grid logical name space"
- Within a "data grid logical name space"
- From one "data grid logical name space" to a federated "data grid logical name space"

Each level of data manipulation requires bulk operations for loading content, registering context, and managing wide area network latency. Examples of bulk operations provided within the Storage Resource Broker data grid include:

- Bulk file registration (aggregation of information about files on a local file system and the bulk loading of the information into the data grid metadata catalog)
- Bulk file loading (aggregation of small files on a local file system into containers before transport, the movement of the containers to a storage system controlled by the data grid, and the bulk registration of the information about the files into the data grid metadata catalog)
- Bulk metadata loading (aggregation of preservation metadata before loading into the preservation environment)
- Client and server initiated parallel I/O streams for data transport (both versions of parallel I/O are needed to handle interactions with fire walls)

Once the electronic records are registered and loaded into a data grid, scalable browsing and discovery mechanisms are then needed to support collections with millions of files. A simple "ls" command to list all the files within a collection will not be adequate. Listing of the items within a collection requires the ability to page through metadata, with a specified number of records displayed on each request, and the ability to specify the desired page. Discovery based on queries of the preservation metadata is scalable through current database technology.

Automation of data and metadata movement between federated data grids requires the ability to cross-register the five name spaces. Each data grid needs to be able to share the set of resource names, the user names, the file names, and the metadata names associated with each file. The cross-registration is done through use of bulk metadata registration operations. The actual replication of data between the data grids requires bulk file manipulations for file registration and file loading. A current development activity for the SRB data grid is the demonstration of bulk replication of files directly between servers located in two independent data grids.

Scalability also requires support for latency management (basically mechanisms that decrease the number of messages sent over wide-area-networks) and for consistency management (mechanisms that synchronize content with context after partially completed operations). Examples include:

- Remote procedure execution to support metadata extraction from files
- Replication synchronization after changes to a file
- File staging to disk from tape-based systems
- Metadata synchronization between federated data grids

Through combinations of these bulk operation, latency management, and consistency capabilities, it is possible to build scalable mechanisms for the application of archival processes in preservation environments.

5. Example implementation

The first four logical name spaces (resources, files, users, metadata) have been implemented in the Storage Resource Broker (SRB) data grid [9]. The consistency and access constraints are currently hard-coded in the SRB software. An NSF Information Technology Research project for the implementation of "Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives" is developing the ability to name and manage constraints as the fifth logical name space. Its implementation will allow the dynamic application of policy and preservation constraints on archived material.

The SRB is used at the San Diego Supercomputer Center to implement preservation environments for the National

Archives and Records Administration [10], the National Historical Publications and Records Commission [11], the NSF National Science Digital Library (NSDL) [12], and the University of California San Diego Libraries. The SRB is also used to implement data grids for sharing data and digital libraries for publishing data. The largest data grid (in terms of number of storage systems) is the NSF National Partnership for Advanced Computational Infrastructure that manages data distributed across 86 storage systems. The largest data grid in geographic extent is the NIH Bio-medical Informatics Research Network [13] which links Grid Bricks installed at 17 sites from the West Coast to the East Coast of the United States. The largest data grid federation links seven data grids for the BELLE high-energy physics experiment at KEK [14] in Japan. The data grids are located in Japan, the Far East, Australia, Poland, and the United States and effectively span the world. Data is replicated between the data grids through cross-registered resource, user, and file name spaces.

The total amount of data under SRB data grid management at SDSC as shown in Table 1 is over 350 Terabytes and over 52 million files, indicating the scalability of the SRB technology.

Table 1. Collections housed at SDSC using the Storage Resource Broker

Storage Resource Broker Collections at SDSC (12/16/2004)	GBs of data stored	Number of files	Number of Users
Data Grid			
NSF/ITR - National Virtual Observatory [15]	53,862	9,536,751	100
NSF - National Partnership for Advanced Computational Infrastructure	22,781	6,018,807	380
Hayden Planetarium - Evolution of the Solar System visualizations	7,201	113,600	178
NSF/NPACI - Joint Center for Structural Genomics [16]	5,340	726,575	50
NSF/NPACI - Biology and Environmental collections	8,575	41,376	67
NSF - TeraGrid, ENZO Cosmology simulations [17]	139,510	1,114,323	3,120
NIH - Biomedical Informatics Research Network [13]	6,948	4,451,271	222
Digital Library			
NLM - Digital Embryo image collection [18]	720	45,365	23
NSF/NPACI - Long Term Ecological Reserve [19]	253	8,885	36
NSF/NPACI - Grid Portal	2,365	51,261	460
NIH - Alliance for Cell Signaling micro-array data [20]	882	65,527	21
NSF - National Science Digital Library SIO Explorer collection [21]	2,104	826,587	27
NSF/NPACI -Transana education research video collection	92	2,387	26
NSF/ITR - Southern California Earthquake Center [22]	95,392	1,877,537	63
Persistent Archive			
UCSD Libraries image collection	128	205,215	29
NARA- Research Prototype Persistent Archive [10]	865	421,038	58
NSF - National Science Digital Library persistent archive [12]	3,572	26,918,638	136
TOTAL	356 TB	52 million	4,996

The number of users with controlled access to selected sub-collections is nearly 5000. The data collections are housed on multiple resources at SDSC including an HPSS archive, a SAM-QFS archive, a Sun SAN file system, Grid Bricks based on commodity disk, and Oracle databases.

Several of the collections use geographically remote storage resources, with data distributed between multiple sites. The most notable facts are that single project collections now are over

100 Terabytes in size, with many collections that are Terabytes to tens of Terabytes in size. Also, the number of files in the SRB collections is much greater than the number that can be effectively managed in a single file system. The ability to manage 50 million files depends strongly on the aggregation of small files into containers, as well as the use of multiple storage systems. The average file size is 6.8 megabytes. If the NSDL collection is excluded (which has an average file size of 132 kilobytes), the average file size of the remaining collections is 14 megabytes. This file size is large enough to use Grid Bricks for permanent on-line access without resorting to containers. For the NSDL collection, the use of containers was essential for the organization and management of the data.

The NARA research prototype persistent archive is implemented through the federation of three SRB data grids located at SDSC, NARA, and the University of Maryland. Figure 1 shows a schematic of the distributed preservation environment.

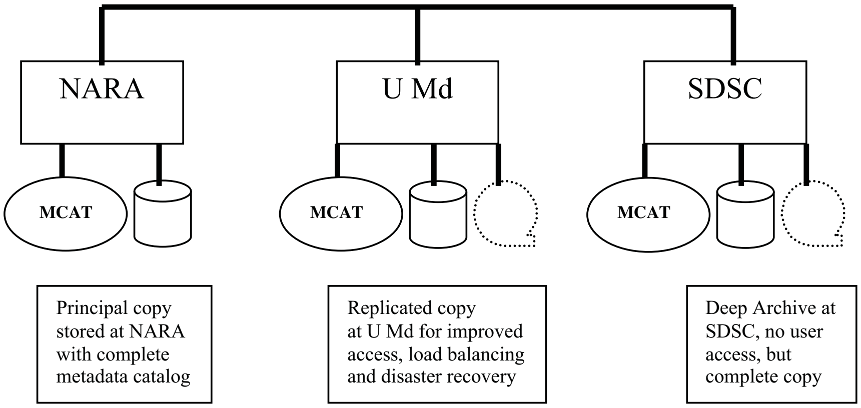
NARA
U Md
SDSC
MCAT
MCAT
MCAT
Principal copy stored at NARA with complete metadata catalog
Replicated copy at U Md for improved access, load balancing and disaster recovery
Deep Archive at SDSC, no user access, but complete copy

Figure 1. Federated Data Grids for Mitigating Risk of Loss of Authenticity

The federation was designed to address each of the types of risk related to data management.

- Media failure – multiple copies are maintained at both SDSC and the University of Maryland. Each site maintains a copy on a file system and on a tape-based archival storage system.
- Vendor-hardware/software systemic error – different vendor systems are used at the three sites. SDSC uses an Oracle database for the metadata catalog, and commodity-based Grid Bricks and a Sun Sam-QFS archive for files. The University of Maryland uses an Informix database for the metadata catalog, a Sun FastT200 disk system and an IBM HPSS archive for files. NARA uses an Oracle database and a Sun FastT200 disk system.
- Operational error – Upgrades to the systems are done at different times, making it possible to validate operational procedures independently at each site. This includes upgrades to the SRB data grid. By providing separate metadata catalog instances for each version of the SRB data grid, it is possible to upgrade and test a new version of a data grid within one site before installation at another site.
- Natural disaster – A common natural disaster (fire, flood, earthquake, hurricane) between SDSC and the University of Maryland is highly improbable.
- Malicious user – By implementing a staging environment for the ingestion of electronic records into the data grid at SDSC, and by

restricting user access, remote interactions with the system can be minimized.

Both data and metadata are replicated between the data grids at NARA and the University of Maryland. Consistency control polices and access control policies are used to control updates to preservation metadata, as well as access to content within a remote data grid.

Multiple collections have been loaded into the environment. The largest collection is a 1.2 terabyte image collection that is being replicated between the University of Maryland and SDSC. The replication is being used to test the ability of federated data grids to support bulk data movement between data grid servers located in different data grids under appropriate consistency and access control constraints.

6. Conclusion

Commodity-based storage systems are suitable for holding on-line copies of preservation data. Data Grid technology makes it possible to implement the uniform name spaces that are needed to manage data distributed across multiple disk file systems and tape archives. Data grids also provide the federation mechanisms needed to replicate authenticity information between multiple databases. The ability to replicate data over wide area networks makes it possible to build a preservation environment that mitigates multiple types of risk of data loss. Preservation environments are being successfully implemented on data grid technology.

7. References

1. Moore, R., A. Merzky, “Persistent Archive Concepts,” Global Grid Forum, December 2003.
2. Moore, R., “Preservation Environments,” NASA / IEEE MSST2004, Twelfth NASA Goddard / Twenty-First IEEE Conference on Mass Storage Systems and Technologies, April 2004
3. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. January 2002. ISO 14721:2003.
4. Rajasekar, A., Michael Wan, Reagan Moore, George Kremenek, Tom Guptil, “Data Grids, Collections, and Grid Bricks”, Proceedings of the 20th IEEE Symposium on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies, San Diego, April 2003.
5. Rajasekar, A., M. Wan, R. Moore, W. Schroeder, “Data Grid Federation”, PDPTA 2004 - Special Session on New Trends in Distributed Data Access, June 2004.
6. Moore, R., R. Marciano, “Prototype Preservation Environments”, submitted to Library Trends, Dec. 2004.
7. Moore, R., C. Baru, A. Rajasekar, R. Marciano, M. Wan: Data Intensive Computing, In ``The Grid: Blueprint for a New Computing Infrastructure'', eds. I. Foster and C. Kesselman. Morgan Kaufmann, San Francisco, 1999.
8. Moore, R., “The San Diego Project: Persistent Objects,” Archivi & Computer, Automazione E Beni Culturali, l’Archivio Storico Comunale di San Miniato, Pisa, Italy, February, 2003.
9. Baru, C., R, Moore, A. Rajasekar, M. Wan,"The SDSC Storage Resource Broker,” Proc. CASCON'98 Conference, Nov.30-Dec.3, 1998, Toronto, Canada.
10. NARA Persistent Archives project, http://www.sdsc.edu/NARA/
11. PAT – Persistent Archive Testbed, http://www.sdsc.edu/PAT
12. NSDL – National Science Digital Library, http://www.nsdl.org/
13. BIRN – The Biomedical Informatics Research Network, http://www.nbirn.net.
14. http://www.kek.jp/intra-e/index.html

15. NVO – National Virtual Observatory, http://www.us-vo.org/
16. JCSG – Joint Center for Structural Genomics, http://www.jcsg.org/
17. ENZO – Cosmological simulation code, http://cosmos.ucsd.edu/enzo/
18. Digital Embryo – collection of images for embryology courses, http://netlab.gmu.edu/visembryo/index.html
19. LTER, US Long Term Ecological Research network, http://lternet.edu/
20. AFCS – Alliance for Cell Signaling, http://www.afcs.org
21. SIO Explorer Digital Library Project to provide education and research material from oceanographic voyages in collaboration with NSDL, http://nsdl.sdsc.edu/.
22. SCEC – Southern California Earthquake Center community digital library, http://www.sdsc.edu/SCEC/